제3강 표본분포

1. 용어 정의
2. 표본평균의 분포
3. 표본분산의 분포
4. 이론적 표본분포

1. 용어 정의

(1) (표본)통계량(sample statistic)

- 모집단으로 부터 추출된 표본의 수량적인 특성을 (표본)통계량이라고 함
- 예 : 표본평균($\bar{X}$), 표본분산(s^2)

(2) 표본분포(sampling distribution)

- (표본)통계량의 확률분포를 표본분포라고 함
- 모수의 값은 일정하게 정해져 있으나 표본통계량의 값은 표본에 따라 바뀌므로 표본통계량은 확률변수로서 일정한 확률분포를 이루고 있음

(3) 확률표본(random sample)

- 확률변수 X가 특정 확률분포를 따른다고 할 때, 이 확률분포로부터 각각 독립적으로 n개의 표본으로 각각의 관찰 값들은 서로 독립이며 동일한 분포를 가짐
- $\boldsymbol{P(X_1) = P(X_2) = \cdots = P(X_n) = P(X)}$ (확률표본의 확률분포가 모집단의 확률분포와 동일)
- $\boldsymbol{P(X_1 X_2 \ldots X_n) = P(X_1)P(X_2) \ldots P(X_n)}$(각 표본이 서로 독립)

(실험) 10,000명의 학생이 있는 J대학교 학생들의 키의 분포가 평균이 168cm이고, 분산이 25cm인 정규분포 N(168,25)를 따른다고 할 때, 학생 10명을 임의로 추출하여 키를 측정

- 확률변수 X가 '학생들의 키'를 나타낸다고 할 때, 표본으로 추출된 10명의 학생의 키는 확률변수로 나타낼 수 있음. 즉, $\{X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}\}$
- X_1이 '처음 추출된 학생의 키'라고 하면 X_1의 분포는 전체 학생의 분포와 같고(즉, $X_1 \sim N(168, 25)$), 이는 X_1이 표본을 반복해서 추출할 때마다 다른 값을 가질 수 있다는 것을 의미함
- 또한 $X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}$에 대해서도 각 값은 표본을 반복해서 추출할 때마다 다른 값을 가질 수 있으므로 $X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}$의 분포는 다음과 같이 나타낼 수 있음

$X_i \sim N(168, 25), i = 1, 2, \ldots, 10$

- 두 개의 다른 표본이 다음의 표와 같이 추출되었다면 <표본 1>에서 $X_1 = 165$이나 <표본 2>에서 $X_1 = 180$임

구분	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
표본 1	165	173	159	183	171	169	165	173	149	177
표본 2	180	144	163	172	185	172	166	174	170	165

- 위의 예에서 가능한 표본의 수는 10,000명에서 10명을 추출하는 방법인 ${}_{10000}C_{10}$과 같으므로 X_1의 가능한 값은 ${}_{10000}C_{10}$개 임.
- 따라서 X_1의 분포는 ${}_{10000}C_{10}$개의 X_1값 분포를 의미하며 이 값의 분포는 모집단의 분포 $X \sim N(168, 25)$와 같음

2. 표본평균의 분포

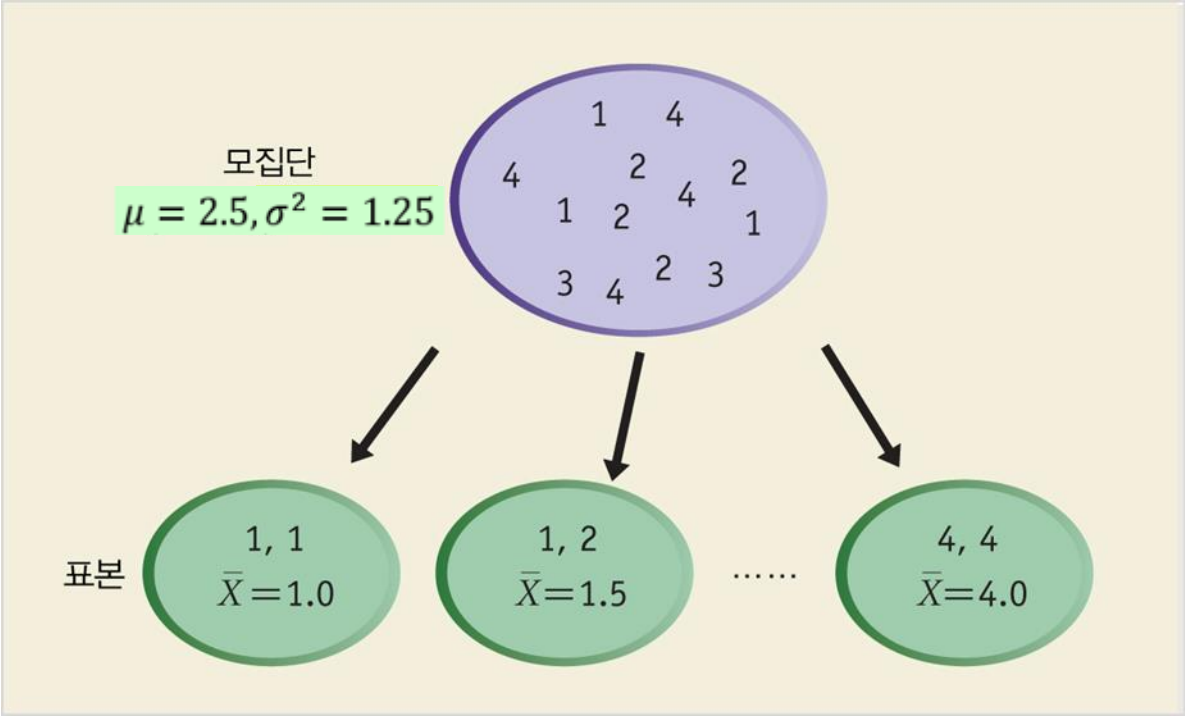
(1) 배경

- 평균 μ와 분산 σ^2를 갖고 정규분포를 하는 모집단에서 추출된 표본평균 $\bar{X}$는 표본이 어떻게 추출되느냐에 따라 다른 표본평균을 가지고 있어 표본평균 그 자체가 확률변수임

(예시) J제약회사는 많은 종류의 신약을 개발하였다. 이 제약회사가 신약을 개발하기 위해서 1,2,3 혹은 4년의 시간이 걸렸으며 각각의 발생확률은 동등하다고 가정하면, 신약의 평균 개발기간인 모집단 평균 및 분산은 다음과 같이 구할 수 있음

- $\mu = E(X) = \sum xf(x) = \left(1 \times \frac{1}{4}\right) + \cdots + \left(4 \times \frac{1}{4}\right) = 2.5$

$\sigma^2 = Var(X) = \sum (x-\mu)^2 P(x) = (1-2.5)^2 \times \frac{1}{4} + \cdots + (4-2.5)^2 \times \frac{1}{4} = 1.25$



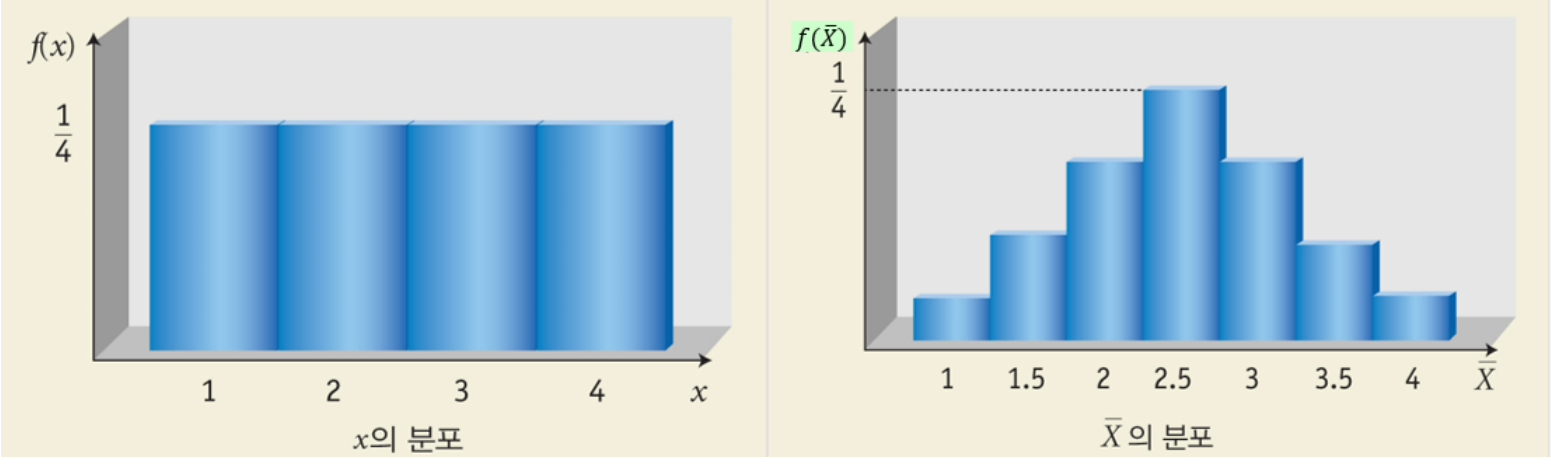
- 표본크기 2(즉, n=2)의 모든 가능한 표본과 표본평균은 다음의 표와 같음

표본	표본평균($\bar{X}$)	표본	표본평균($\bar{X}$)
1, 1	1.0	3, 1	2.0
1, 2	1.5	3, 2	2.5
1, 3	2.0	3, 3	3.0
1, 4	2.5	3, 4	3.5
2, 1	1.5	4, 1	2.5
2, 2	2.0	4, 2	3.0
2, 3	2.5	4, 3	3.5
2, 4	3.0	4, 4	4.0

- 표본평균의 확률분포는 다음의 표와 같음

$\bar{X}$	1	1.5	2	2.5	3	3.5	4
$P(\bar{X})$	1/16	2/16	3/16	4/16	3/16	2/16	1/16

- 확률변수 x 및 표본평균 $\bar{X}$의 분포는 다음의 그림과 같음

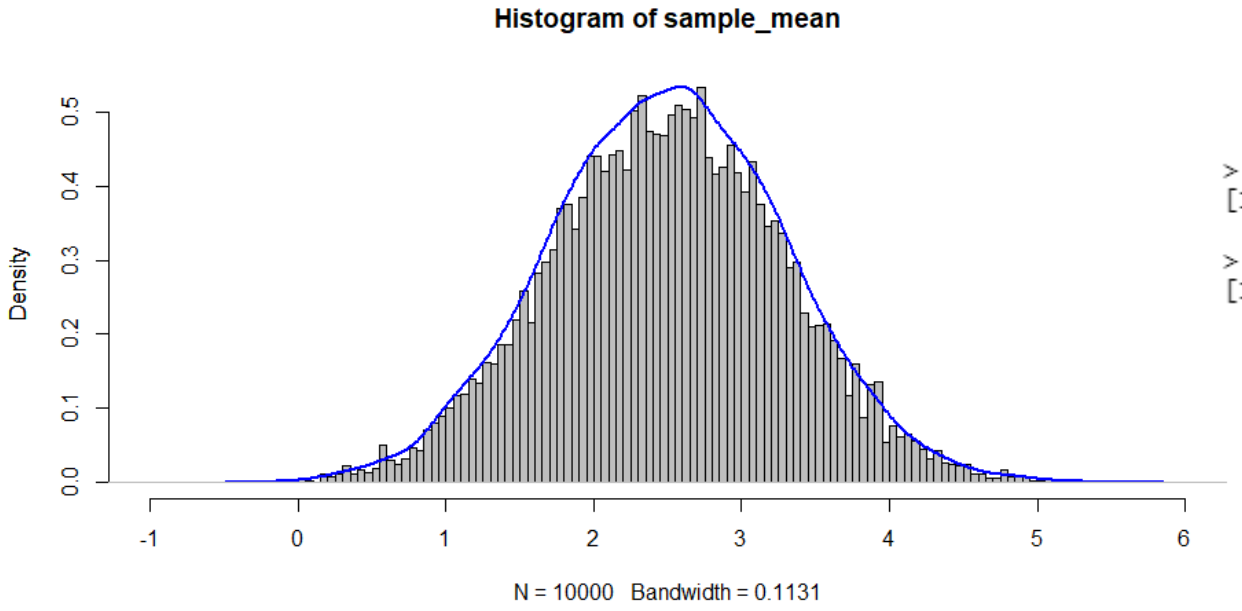


– $\bar{X}$의 평균과 분산은 각각 다음과 같음

· 평균 : $\mu_{\bar{X}} = E(\bar{X}) = \sum \bar{X} P(\bar{X}) = \left(1.0 \times \frac{1}{16}\right) + \left(1.5 \times \frac{2}{16}\right) + \cdots + \left(4.0 \times \frac{1}{16}\right) = 2.5$

· 분산 : $\sigma_{\bar{X}}^2 = Var(\bar{X}) = \sum (\bar{X} - \mu_{\bar{X}})^2 P(\bar{X}) = (1.0 - 2.5)^2 \times \frac{1}{16} + (1.5 - 2.5)^2 \times \frac{2}{16} + (4.0 - 2.5)^2 \times \frac{1}{16} = 0.625$

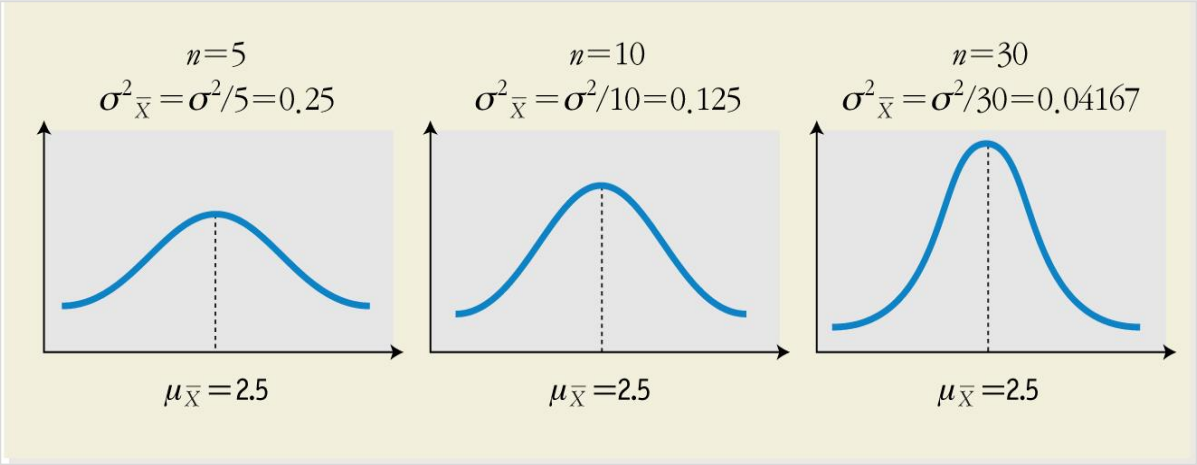
– 계산결과를 보면 두 분포의 평균은 2.5로 같지만 x의 분산 σ_X^2과 $\bar{X}$의 분산 $\sigma_{\bar{X}}^2$은 서로 같지 않으며, $\sigma_{\bar{X}}^2$은 σ_X^2의 $\frac{1}{2}$임.

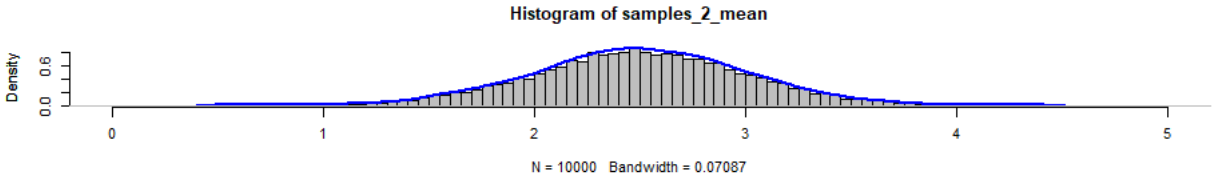



```
> (mean(sample_mean))
[1] 2.494004

> (var(sample_mean))
[1] 0.6284706
```

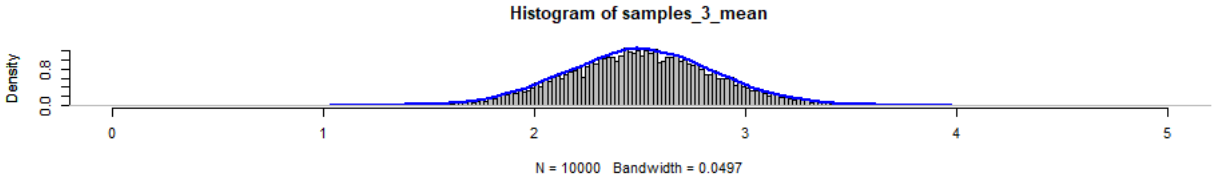
- n=5,10,30일 경우 표본평균 $\bar{X}$의 표본분포는 다음의 그림과 같음





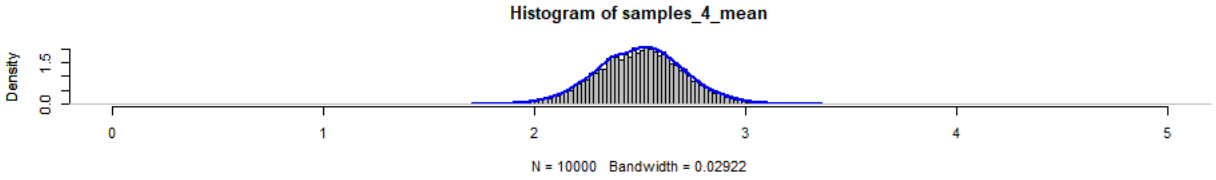

```
> mean(samples_2_mean)
[1] 2.49543

> var(samples_2_mean)
[1] 0.2497486
```




```
> mean(samples_3_mean)
[1] 2.502499

> var(samples_3_mean)
[1] 0.122446
```




```
> mean(samples_4_mean)
[1] 2.499477

> var(samples_4_mean)
[1] 0.04197086
```

- 결론 : 표본평균의 평균은 모집단의 평균과 같고, 표본평균의 분산은 모집단의 분산을 n으로 나눈 것과 같음

(2) 표본평균의 표본분포 특성

– 평균 $\boldsymbol{\mu}$ 와 분산 $\boldsymbol{\sigma^2}$을 갖는 모집단으로부터 크기 n의 임의표본을 추출하였으며, 이 표본의 평균을 $\bar{\boldsymbol{X}}$라고 하면,

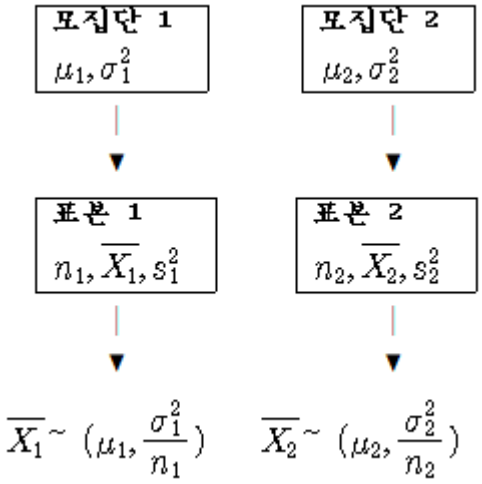
- $\mu_{\bar{X}} = \mu$
- $\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}$
- $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$ ($\bar{\boldsymbol{X}}$의 표준편차는 평균의 표준오차(standard error)라고도 함)

– 대수의 법칙(law of large number) : 표본의 크기(n)가 클수록 표본평균 $\bar{\boldsymbol{X}}$가 모평균 $\boldsymbol{\mu}$에 가까워질 확률이 크게 된다. 따라서 $\boldsymbol{X_1, X_2, X_3, \ldots, X_n}$이 평균이 $\boldsymbol{\mu}$인 확률표본일 때 표본크기 n이 커짐에 따라 $\bar{\boldsymbol{X}}$는 $\boldsymbol{\mu}$에 접근한다.

(3) 두 표본평균 차이의 분포

– 두 표본평균(서로 독립이라고 가정)의 차이 즉, $\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2$의 평균과 분산은 다음과 같음

- $\boldsymbol{E(\bar{X}_1 - \bar{X}_2) = E(\bar{X}_1) - E(\bar{X}_2) = \mu_1 - \mu_2}$
- $\boldsymbol{Var(\bar{X}_1 - \bar{X}_2) = Var(\bar{X}_1) - Var(\bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$



(4) 중심극한정리

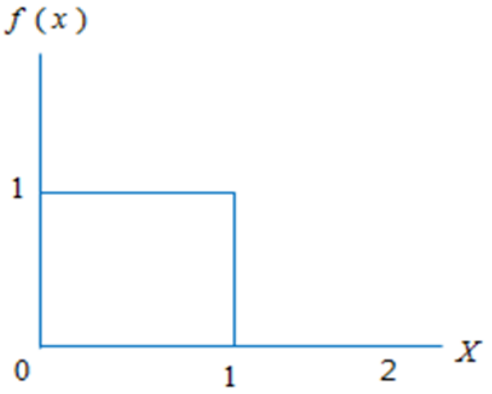
① 배경

- 관측치의 모집단이 정규분포에 따를 경우 관측치 역시 정규분포에 따르고, 관측치의 모집단이 실제로 정규분포가 아니면 관측치 역시 정규분포에 따르지 않음
- 그러나 관측치의 모집단이 실제로 정규분포가 아닌 경우에도 중심극한정리에 의해 정규확률분포를 이용한 추정량의 근사확률을 구할 수 있음.

(예1) 확률변수 X가 균등분포 $U(0,1)$을 따른다고 할 때 $X \sim U(0,1)$이며, X는 0과 1사이에서 균등한 분포를 갖는 연속확률변수인데 이러한 확률분포로부터 크기 n=11인 확률표본을 1,000개 추출하는 실험.

- 1,000개 표본 중 처음 5개의 표본이 다음과 같다고 하면, 각 표본의 평균 $\bar{x}_1, \bar{x}_2, .., \bar{x}_5$를 구할 수 있음.

표본	관측값											$\bar{X}$
1	0.316	0.377	0.357	0.548	0.564	0.312	0.217	0.373	0.535	0.963	0.746	0.482
2	0.481	0.611	0.478	0.512	0.611	0.841	0.152	0.956	0.604	0.825	0.402	0.589
3	0.233	0.724	0.231	0.779	0.077	0.839	0.779	0.487	0.140	0.498	0.221	0.455
4	0.907	0.149	0.624	0.609	0.291	0.398	0.596	0.371	0.929	0.761	0.949	0.599
5	0.809	0.055	0.299	0.466	0.625	0.088	0.958	0.742	0.955	0.885	0.598	0.589



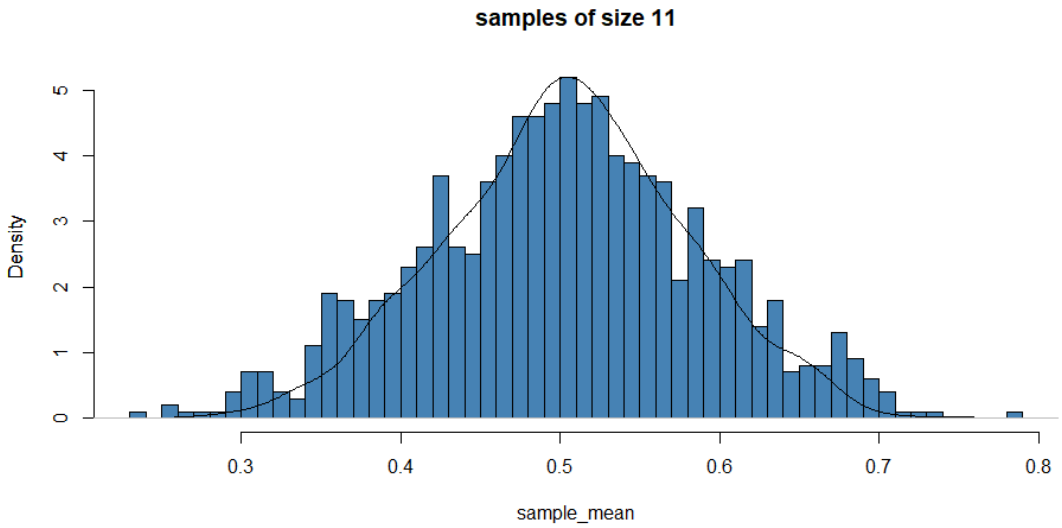
- 1,000개 표본으로부터 구한 표본평균 $\bar{x}_1, \bar{x}_2, .., \bar{x}_{1000}$을 이용하여 막대그림표를 그리면 다음과 같음.
- 즉, 균등분포로부터 구한 표본평균의 분포가 정규분포와 근사한 분포를 가짐
- $X \sim U(0,1)$에서 $E(X) = \frac{1}{2}, Var(X) = \frac{1}{12}$이므로 n=11인 경우 표본평균 $\bar{X}$의 평균과 분산은 다음과 같음

· $E(\bar{X}) = \frac{1}{n} nE(X) = E(X) = \frac{1}{2}$

· $Var(\bar{X}) = \frac{1}{n^2} nVar(X) = \frac{1}{n} Var(X) = \frac{1}{11} \times \frac{1}{12} = \frac{1}{132} = 0.008$

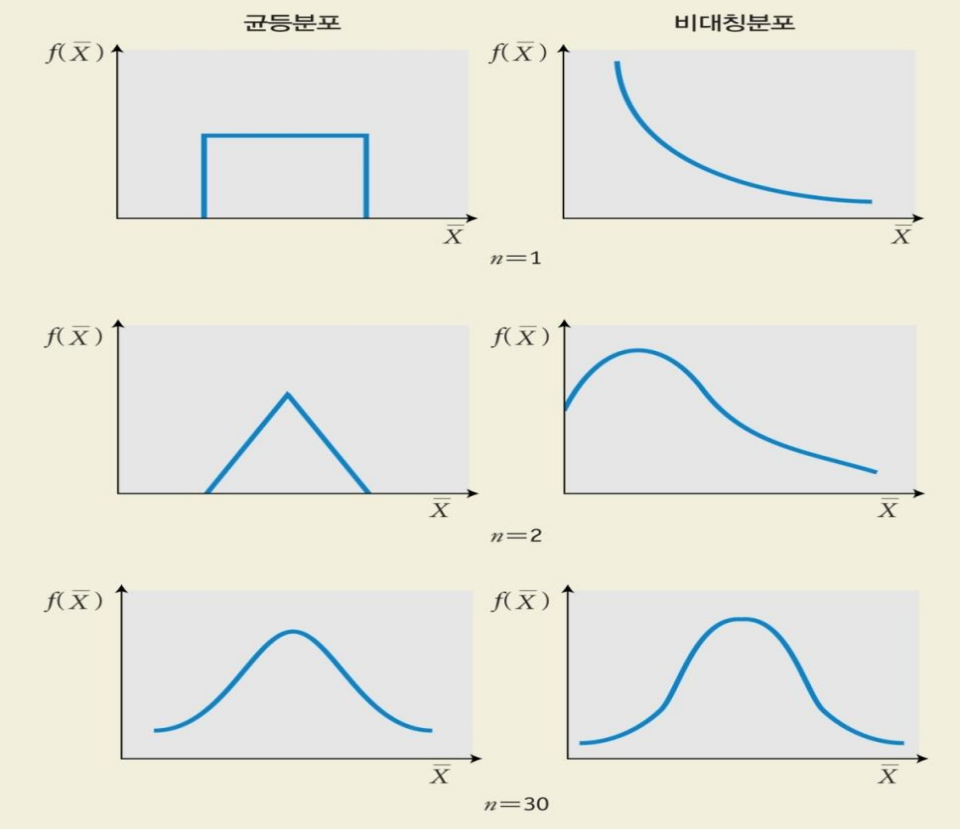
· $\sqrt{Var(\bar{X})} = \sqrt{0.008} = 0.087$

- 즉, 표본평균 $\bar{X}$는 근사적으로 $N(\frac{1}{2}, \frac{1}{132})$을 따름
- 1,000개 표본의 표본평균의 평균과 표준편차는 각각 0.5019과 0.0078로서 모집단의 이론적인 평균과 표준편차에 근접함을 알 수 있음

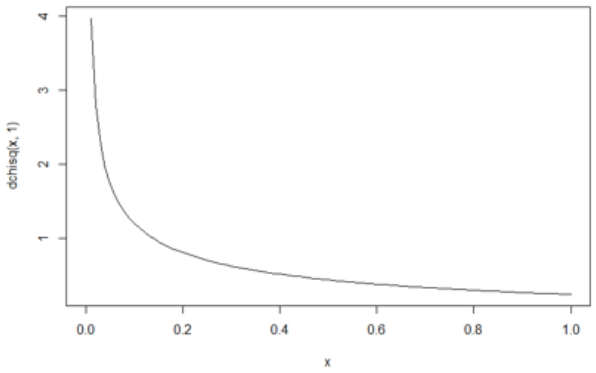


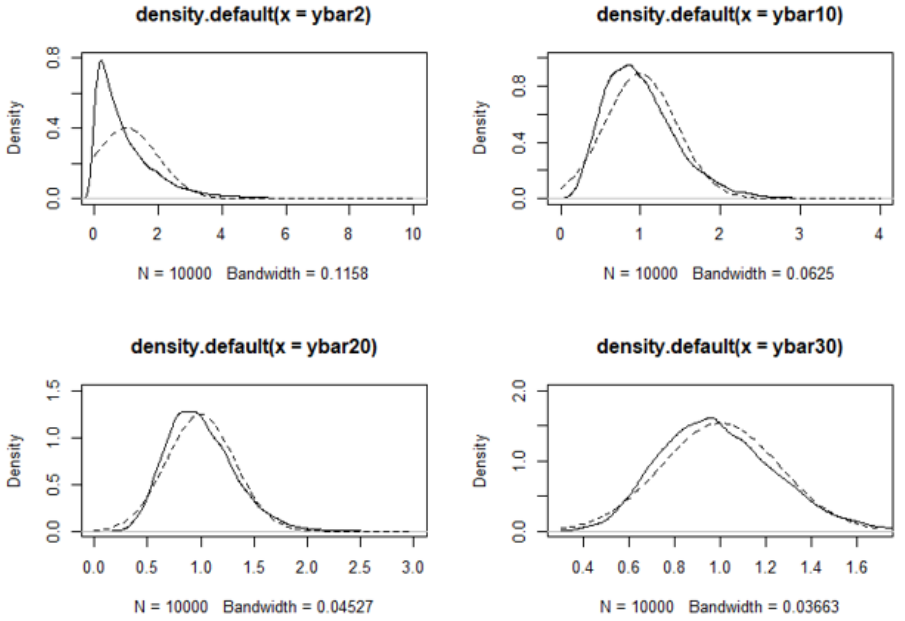

```
> (mean(sample.avgs))
[1] 0.5019778
> (sd(sample.avgs))
[1] 0.08852728
> (var(sample.avgs))
[1] 0.007837079
```

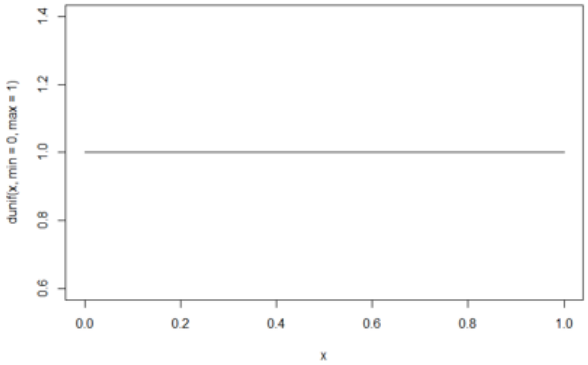
- n의 변화에 따른 $\bar{X}$의 확률밀도함수의 모양은 다음의 그림과 같음

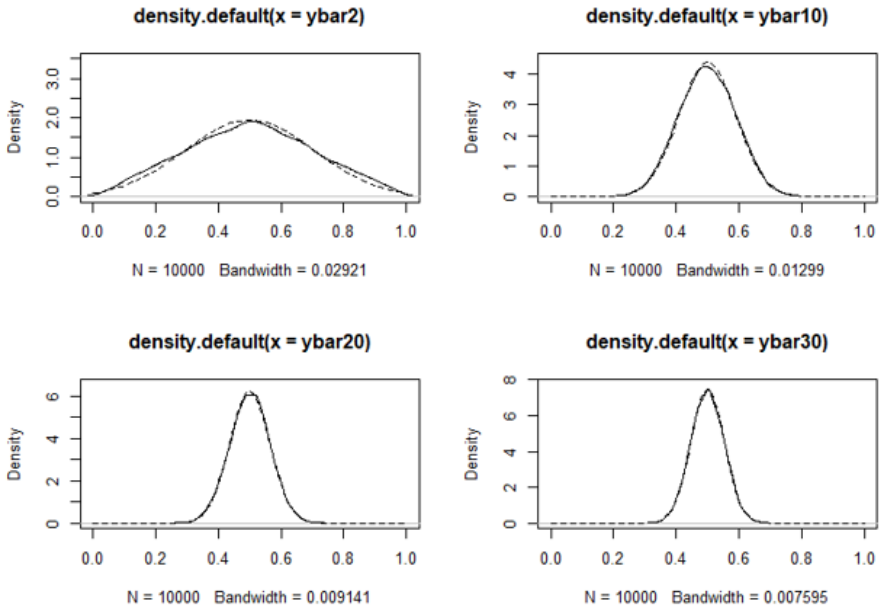


- n의 변화에 따른 $\bar{X}$의 확률밀도함수의 모양은 다음의 그림과 같음









② 중심극한정리(Central Limit Theorem)

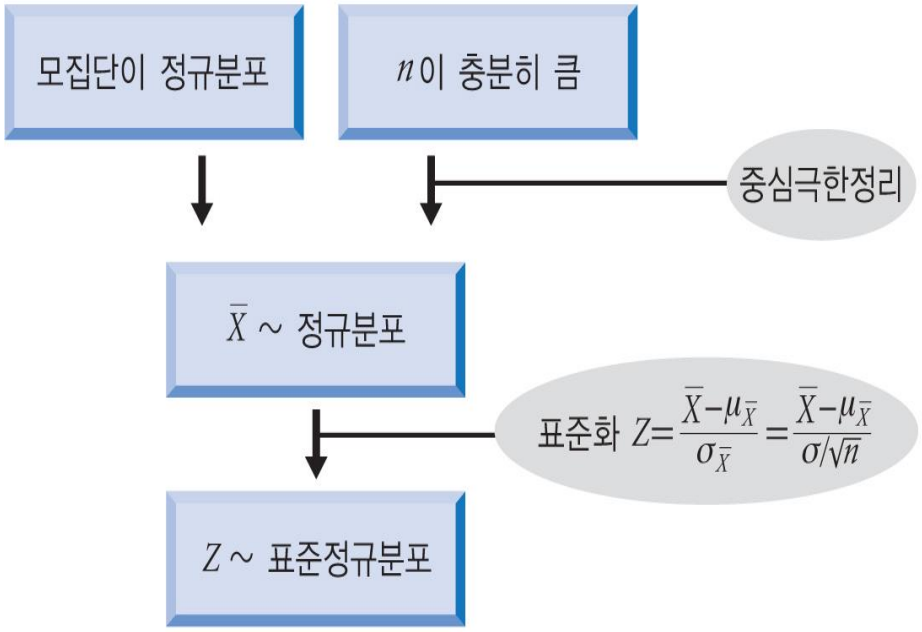
- 만약에 확률표본 $X_1, X_2, X_3, \ldots, X_n$이 평균 μ와 분산 σ^2을 갖는 정규분포에서 추출되었다면, 표본평균 $\bar{X}$의 분포는 표본의 크기 n에 관계없이 평균 μ와 분산 $\frac{\sigma^2}{n}$을 갖는 정규분포를 따름. 즉, $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$

- 평균 μ와 분산 σ^2을 갖는 갖는 확률변수 X가 비록 정규분포가 아닌 다른 어떤 분포를 하더라도 표본평균 $\bar{X}$는 표본의 크기 n이 커짐에 따라 평균 μ와 분산 $\frac{\sigma^2}{n}$을 갖고 정규분포에 접근하는데 이를 중심극한정리라고 함. 즉, n을 크게 하면 $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$

(5) 표본평균의 표준화

- 평균이 μ이고 분산이 σ^2인 확률변수의 X를 표준화하는 방법은 평균을 빼주고 표준편차로 나눔

- 따라서 평균이 μ이고 분산이 $\frac{\sigma^2}{n}$인 표본평균 $\bar{X}$의 표준화된 확률변수는 $\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$ 임



3. 표본분산의 분포

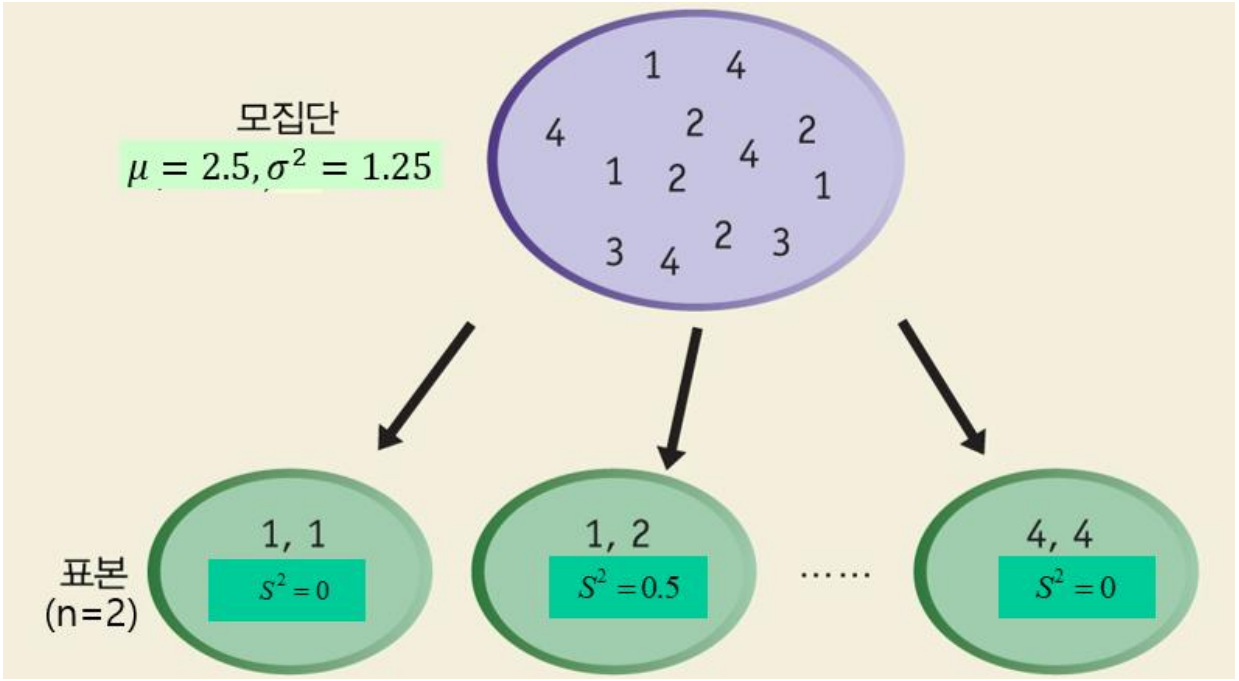
(1) 배경

- 평균 μ와 분산 σ^2를 갖고 정규분포를 하는 모집단에서 추출된 표본평균 $\bar{X}$는 표본이 어떻게 추출되느냐에 따라 다른 표본평균을 가지고 있어 표본평균 그 자체가 확률변수임

(예 2) J제약회사는 많은 종류의 신약을 개발하였다. 이 제약회사가 신약을 개발하기 위해서 1,2,3 혹은 4년의 시간이 걸렸으며 각각의 발생확률은 동등하다고 가정하면, 신약의 평균 개발기간인 모집단 평균 및 분산은 다음과 같이 구할 수 있음

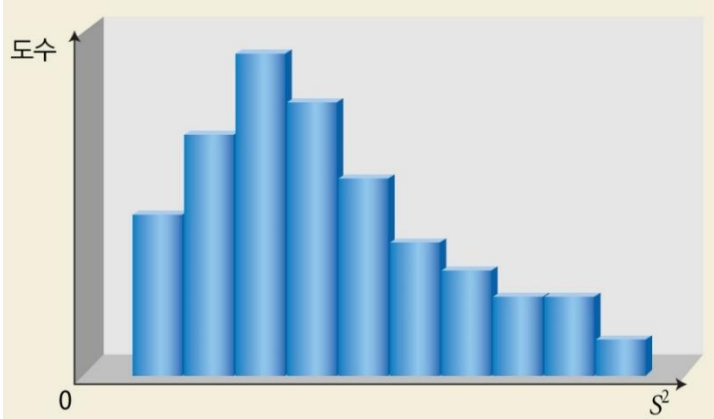
- $\mu = E(X) = \sum xf(x) = \left(1 \times \frac{1}{4}\right) + \cdots + \left(4 \times \frac{1}{4}\right) = 2.5$

$$\sigma^2 = Var(X) = \sum (x-\mu)^2 P(x) = (1-2.5)^2 \times \frac{1}{4} + \cdots + (4-2.5)^2 \times \frac{1}{4} = 1.25$$



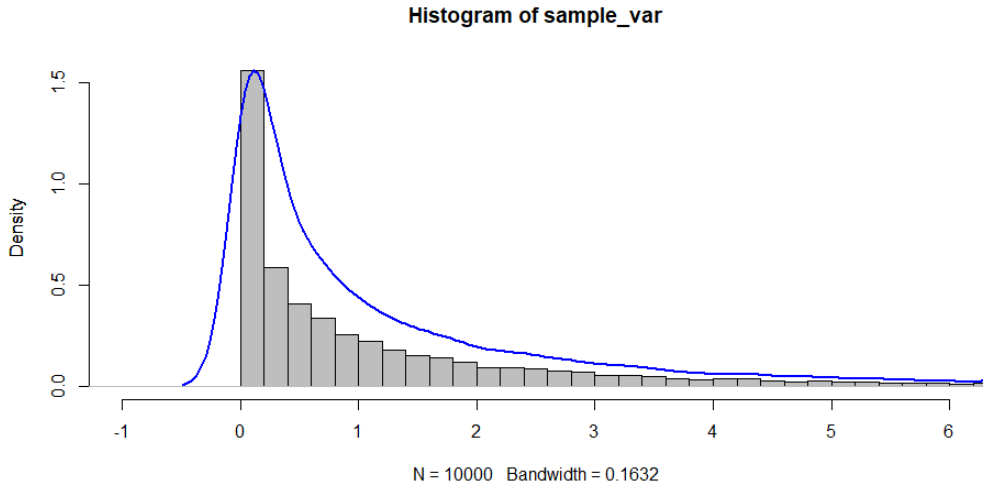
– 표본크기 2(즉, n=2)의 모든 가능한 표본과 표본평균 및 표본분산은 다음의 표와 같고, 표본분산 s^2의 분포는 다음의 그림과 같음

표본	표본평균	표본분산	표본	표본평균	표본분산
(1,1)	1	0	(3,1)	2	2
(1,2)	1.5	0.5	(3,2)	2.5	0.5
(1,3)	2	2	(3,3)	3	0
(1,4)	2.5	4.5	(3,4)	3.5	0.5
(2,1)	1.5	0.5	(4,1)	2.5	4.5
(2,2)	2	0	(4,2)	3	2
(2,3)	2.5	0.5	(4,3)	3.5	0.5
(2,4)	3	2	(4,4)	4	0



(예 3) 정규분포에 따르는 모집단($\mu = 2.5, \sigma = 1.118$) 에서 크기 n=2인 확률표본을 1,000개 추출하는 실험.

– 표본분산의 분포는 정규분포를 따르지 않고 χ^2–분포와 근사한 분포를 가짐.



(예 4) 정규분포에 따르는 모집단($\mu = 10, \sigma = 2$)에서 표본크기 n=11인 확률표본을 1,000개 추출하는 실험.

- 1,000개 표본 중 처음 5개의 표본이 다음과 같으면 각 표본의 표본분산 $s_1^2, s_2^2, .., s_5^2$를 구할 수 있듯이, 정규분포로부터 구한 표본분산 $s_1^2, s_2^2, .., s_{1000}^2$의 분포는 χ^2-분포와 근사한 분포를 가짐

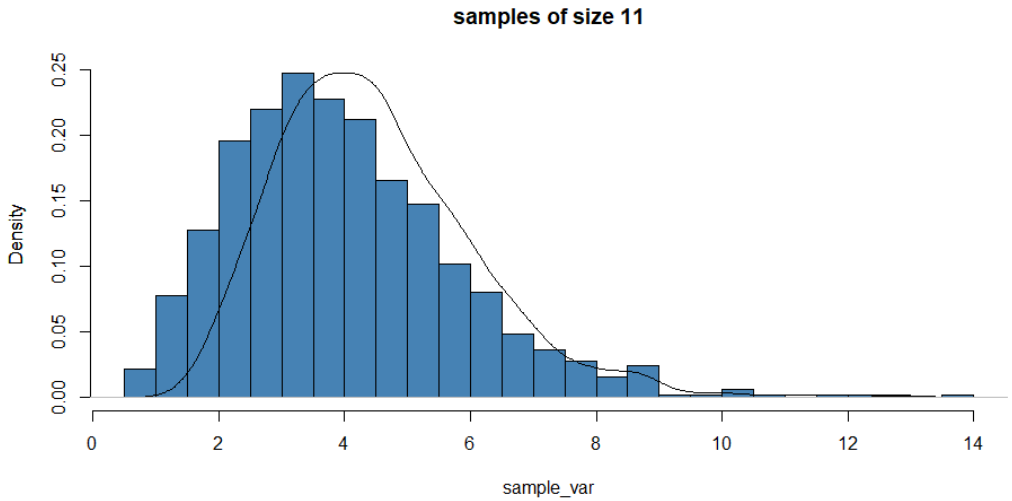
· $E(s^2) = \sigma^2 = 4$

· $Var(s^2) = \frac{2\sigma^4}{n-1} = \frac{2\times 16}{10} = 3.2$

표본	관측값											s^2
1	9.040	9.269	10.320	8.434	10.173	11.325	10.566	10.062	11.994	13.417	11.867	2.146
2	8.541	8.526	7.155	11.538	7.843	8.464	7.918	10.553	9.485	9.342	11.417	2.139
3	11.747	8.945	10.636	13.457	13.399	10.499	7.539	7.731	13.615	9.102	9.370	5.048
4	13.943	10.040	9.544	13.875	10.329	10.425	9.672	8.376	8.382	13.057	11.213	4.011
5	11.896	10.571	8.811	6.546	9.141	14.826	12.580	10.986	9.901	7.635	7.754	5.953

- 즉, 표본분산 s^2은 근사적으로 $\chi^2(4, 3.2)$를 따름

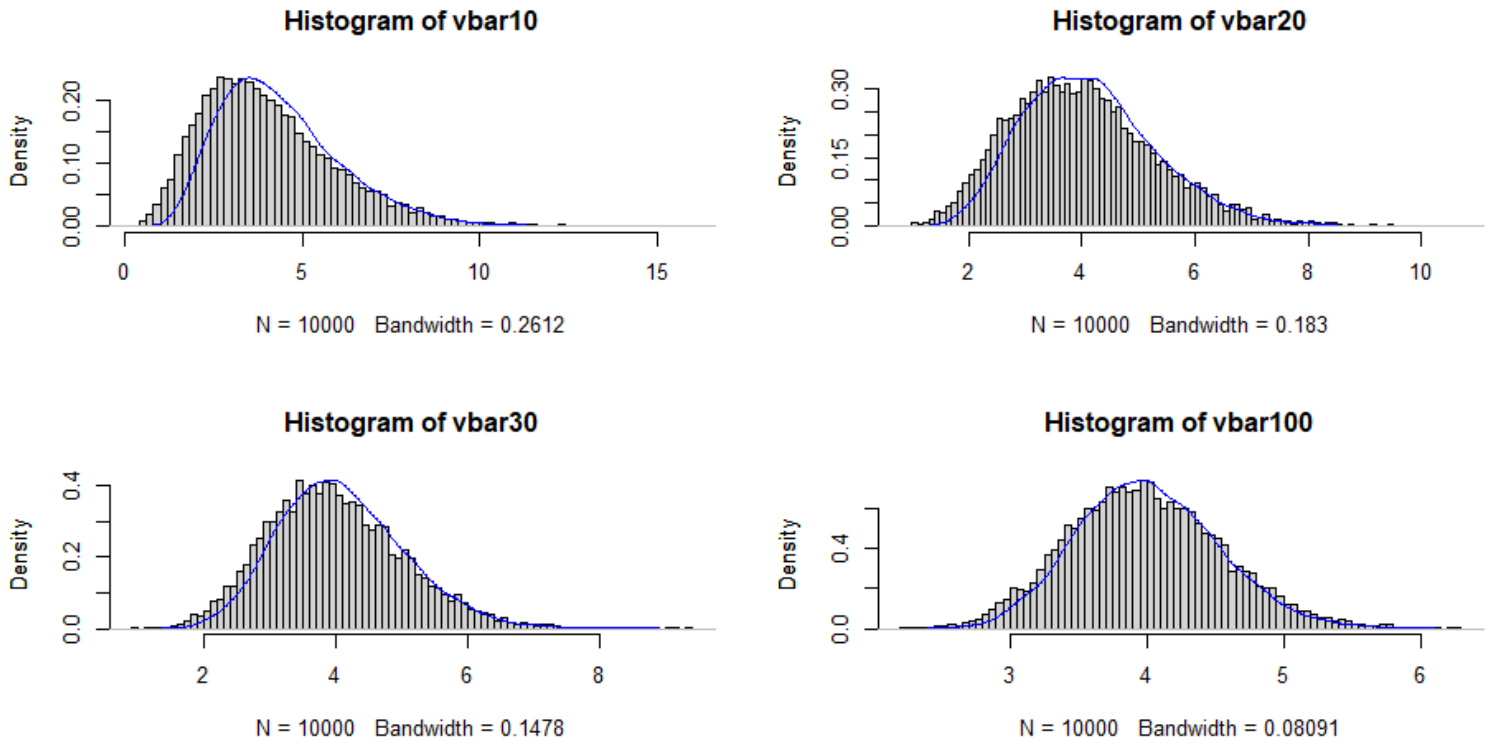
- 1,000개 표본의 표본분산의 평균과 분산은 각각 3.986259 및 3.31678로서 모집단의 이론적인 평균과 분산에 근접함을 알 수 있음




```
> (mean(sample_var))
[1] 3.986259

> (var(sample_var))
[1] 3.31678
```

– 정규분포로부터 구한 표본분산 s^2의 분포는 정규분포에 따르지 않고, 표본의 크기에 따라 모집단의 이론적인 평균 σ^2과 분산 $\frac{2\sigma^4}{n-1}$인 χ^2 –분포를 따름



– 분산 σ^2을 갖는 모집단으로부터 크기 n의 임의표본이 추출되었으며 이 표본의 분산이 s^2이라고 할 때, 표본분산의 평균과 분산은 다음과 같음

· $E(s^2) = \sigma^2 \because E\left[\frac{(n-1)s^2}{\sigma^2}\right] = \frac{(n-1)}{\sigma^2}E(s^2) = n-1$

· $Var(s^2) = \frac{2\sigma^4}{n-1} \because Var\left[\frac{(n-1)s^2}{\sigma^2}\right] = \frac{(n-1)^2}{\sigma^4}Var(s^2) = 2(n-1)$

4. 이론적 표본분포

① χ^2 −분포(chi-square distribution)

- 확률변수 $Z_1, Z_2, \ldots, Z_n$이 서로 독립적으로 표준정규분포 $Z_i \sim N(0,1)$을 따를 때, $Z_1, Z_2, \ldots, Z_n$의 제곱합 $\sum_{i=1}^{n} Z_i^2$은 자유도가 n인 χ^2 −분포를 따름
- $X \sim \chi_n^2$일 때
- $E(X) = n, Var(X) = 2n$
- 분산이 σ^2인 정규분포를 이루는 모집단으로부터 표본크기가 n인 선택 가능한 모든 임의표본이 추출되었을 때, 각 표본의 분산을 s^2이라고 하면 $\frac{(n-1)s^2}{\sigma^2}$은 자유도가 n−1인 χ^2−분포를 따름

즉, $\chi_{n-1}^2 = \frac{(n-1)s^2}{\sigma^2} \sim \chi_{n-1}^2$

- $E(\chi_{n-1}^2) = n-1$
- $Var(\chi_{n-1}^2) = 2(n-1)$

(참고)

$X_i = \mu + e_i, e_i \sim (0, \sigma^2)$ (모집단)

$X_i = \bar{X} + \hat{e}_i, \hat{e}_i \sim (0, s^2)$ (표본)

$\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma^2} = \frac{\sum_{i=1}^{n} e_i^2}{\sigma^2} = \sim \chi_n^2$ (σ^2 및 μ를 알 때)

$\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{\sigma^2} = \frac{\sum_{i=1}^{n} \hat{e}_i^2}{\sigma^2} = \frac{(n-1)s^2}{\sigma^2} \sim \chi_{n-1}^2$ (σ^2은 알고, μ를 모를 때)

② **t** –분포(Student's t-distribution)

– $\boldsymbol{Z \sim N(0,1)}$, $\boldsymbol{V \sim \chi^2_v}$이고 Z와 V가 독립이면, $\boldsymbol{T = \frac{Z}{\sqrt{\frac{V}{v}}} \sim t_v}$임

– $T = \frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}} \sim t_{n-1}$ ⇦ $T = \frac{Z}{\sqrt{\frac{V}{v}}} = \frac{\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}}{\frac{\sqrt{\frac{(n-1)s^2}{\sigma^2}}}{\sqrt{(n-1)}}} = \frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}}$

– 자유도가 무한히 커지면 t-분포는 표준정규분포에 접근

③ **F** –분포(Snedecor's F-distribution)

– $\boldsymbol{X_1 \sim \chi^2_{v_1}}$, $\boldsymbol{X_2 \sim \chi^2_{v_2}}$이고 $\boldsymbol{X_1}$, $\boldsymbol{X_2}$이 서로 독립이면, $\boldsymbol{F = \frac{\frac{X_1}{v_1}}{\frac{X_2}{v_2}} \sim F\binom{v_1}{v_2}}$임

– $F = \frac{\frac{\chi^2_{n_1-1}}{n_1-1}}{\frac{\chi^2_{n_2-1}}{n_2-1}} = \frac{\frac{(n_1-1)s_1^2}{(n_1-1)\sigma_1^2}}{\frac{(n_2-1)s_2^2}{(n_2-1)\sigma_2^2}} = \frac{\frac{s_1^2}{\sigma_1^2}}{\frac{s_2^2}{\sigma_2^2}} \sim F\binom{v_1}{v_2}$ 단, $v_1 = n_1 - 1$, $v_2 = n_2 - 1$